Diversity and population-genetic properties of copy number variations and multicopy genes in cattle

Derek M. Bickhart1,†, Lingyang Xu1,2,†, Jana L. Hutchison1, John B. Cole1, Daniel J. Null1, Steven G. Schroeder1, Jiuzhou Song2, Jose Fernando Garcia3, Tad S. Sonstegard1, Curtis P. Van Tassell1, Robert D. Schnabel4,5, Jeremy F. Taylor4, Harris A. Lewin6, and George E. Liu1,*

1USDA-ARS, Animal Genomics and Improvement Laboratory, Beltsville, MD 20705, USA, 2Department of Animal and Avian Sciences, University of Maryland, College Park, MD 20742, USA, 3Universidade Estadual Paulista (UNESP), Araçatuba, São Paulo, Brazil, 4Division of Animal Sciences, University of Missouri, Columbia, MO 65211, USA, 5Informatics Institute, University of Missouri, Columbia, MO, USA, and 6Department of Evolution and Ecology, University of California, Davis, CA 95616, USA

*(To whom correspondence should be addressed. Tel. +1 301-504-8592. Fax. +1 301-504-8092. E-mail: derek.bickhart@ars.usda.gov (D.M.B.); Tel. +1 301-504-9843. Fax. +1 301-504-8414. E-mail: george.liu@ars.usda.gov (G.E.L.)

†These authors are co-first authors.

Received 5 November 2015; Accepted 29 February 2016

Abstract

The diversity and population genetics of copy number variation (CNV) in domesticated animals are not well understood. In this study, we analysed 75 genomes of major taurine and indicine cattle breeds (including Angus, Brahman, Gir, Holstein, Jersey, Limousin, Nelore, and Romagnola), sequenced to 11-fold coverage to identify 1,853 non-redundant CNV regions. Supported by high validation rates in array comparative genomic hybridization (CGH) and qPCR experiments, these CNV regions accounted for 3.1% (87.5 Mb) of the cattle reference genome, representing a significant increase over previous estimates of the area of the genome that is copy number variable (~2%). Further population genetics and evolutionary genomics analyses based on these CNVs revealed the population structures of the cattle taurine and indicine breeds and uncovered potential diversely selected CNVs near important functional genes, including AOX1, ASZ1, GAT, GLYAT, and KRTAP9-1. Additionally, 121 CNV gene regions were found to be either breed specific or differentially variable across breeds, such as RICTOR in dairy breeds and PNPLA3 in beef breeds. In contrast, clusters of the PRP and PAG genes were found to be duplicated in all sequenced animals, suggesting that subfunctionalization, neofunctionalization, or overdominance play roles in diversifying those fertility-related genes. These CNV results provide a new glimpse into the diverse selection histories of cattle breeds and a basis for correlating structural variation with complex traits in the future.

Key words: cattle genome, population sequencing, copy number variation, taurine, indicine
1. Introduction

Copy number variations (CNVs) are deletions and insertions of genomic sequence between two individuals of a species. Substantial progress has been made in understanding the impacts of CNVs on both normal phenotypic variability and disease susceptibility in human. The majority of previous studies of CNV in domesticated animals have been based on array comparative genomic hybridization (CGH) experiments or single-nucleotide polymorphism (SNP) arrays. As in human, sequence-based approaches are becoming popular for the study of CNVs in domesticated animals.

The CNV distribution within and among species seems to be shaped by mutation, selection, and demographic history. However, unlike SNPs and microsatellites, the population genetics of CNV is largely unknown. Several studies have explored the evolution and adaptation aspects of CNVs in human, and other species. Only a few cases of recent positive selection were found near AMY1, APOBEC3, MAPT, MIF, and UGT2B17 within human populations. While most of simple deletions and simple duplications (75–90%) display extensive linkage disequilibrium (LD) with SNPs, the properties of CNVs not tagged by SNPs remain unexplored in cattle. This is mainly due to the difficulty of genotyping CNV and to the limited sample size in the published CNV studies.

As one of the most important farm animals, cattle are used for a variety of purposes including dairy, beef, leather, and labour. The majority of the global cattle population can be classified into one of two subspecies: humphless (taurine, Bos taurus taurus) and humped (indicine or zebu, Bos taurus indicus) cattle with dramatic phenotypic differences between them. Earlier studies indicated that these two subspecies diverged from the last common ancestor between 0.6 and 2 million yrs ago. There appears to have been two separate domestication events, with taurine cattle likely being domesticated in the Fertile Crescent ~8,000–10,000 yrs ago and indicine cattle in the Indus Valley ~6,000–8,000 yrs ago. A third independent domestication was proposed in Africa; however, a recent study did not support this hypothesis. Since the early 1800s, breed development has primarily been based on phenotypic selection on coat colour and polled phenotypes. More recently, the adoption of effective genetic selection programs and widespread use of artificial insemination resulted in bottlenecked followed by breed expansion. During the last 50 yrs, animal breeding based on quantitative genetics has resulted in remarkable progress in improving milk and meat production traits. Therefore, selective (natural and human-imposed) and non-selective (demographic events and introgression) have driven changes within the cattle genome. Their combined effects have created exceptional phenotypic diversity and genetic adaptation to local environments across the globe within the modern cattle breeds. For example, indicine cattle are better adapted to warm climates and demonstrate superior resistance to tick infestation than do taurine breeds. Likewise, beef and dairy cattle breeds display distinct patterns in selected metabolic pathways related to muscling, marbling, and milk composition traits. Although cattle genome evolution and demographic history have been explored from multiple aspects, the diversity and population genetic properties of CNV in cattle are still unexplored.

In this study, we compare the diversity and population-genetic properties of CNVs in ~70 cattle individuals sequenced to medium coverage (mean ~11.8x). The data set includes multiple individuals from eight representative cattle breeds, representing both the major taurine and indicine breeds used for both dairy and beef purposes. It provides unprecedented genome-wide resolution to interrogate CNV and a unique opportunity to fully explore the population-genetic properties and evolutionary contributions of multicyclic genes related to breed-specific traits.

2. Material and methods

2.1. CNV calling, distribution, and association with other genomic features

We used a previously described segmentation algorithm to call CNVs. Detailed sample selection and CNV calling methods can be found in Supplementary Material online. Association between CNVs and SDs was tested by Spearman’s rank correlation using 100 kb windows as previously described. Additional genomic features were obtained from public databases. Determination of the overlap between CNVRs and genomic features was performed as previously described.

2.2. Population-genetic analyses and heatmap analysis

Inbreeding is a common feature in livestock due to selective mating and widespread use of artificial insemination. We filtered our samples based on known pedigrees constraining Wright’s coefficient of relationship (r) to <0.25 to identify 69 unrelated individuals for the population-genetic analyses.

During the CNV discovery phase, a total of 1,148,528 windows of 1 kb were identified across the whole genome. Population-specific CNVs were estimated using the statistic VST developed by Redon et al. VST is calculated by considering (VT – V3)/VT, where VT is the variance in normalized copy numbers among all unrelated individuals and V3 is the average variance within each population, weighted for population size.

We next selected the top 1% diverse 1 kb windows (n = 80) from the distinct CNV regions to perform CNV genotyping using the partitioning around medoids (PAM) function in R. A PAM procedure was used to cluster copy number ratios into discrete CN genotypes. Similarly, we partitioned the copy numbers of each 1 kb window into three clusters, representing the low, mid, and high ranges and then coded them using the 0, 1, and 2 matrix for SNP genotyping. Population clustering was then performed using STRUCTURE v2.3.3 assuming three ancestral populations (k = 3). This analysis between taurine and indicine cattle was initially run for values of the number of clusters (k) between 2 and 8. Each analysis was performed using 100,000 replicates and 100,000 burn-in cycles under admixture and correlated allele frequencies models. Reynolds’ genetic distances among breeds were calculated using PHYLIP 3.69. To provide statistical support for the resulting clades, 10,000 bootstrap simulations were performed. The phylogenetic trees were visualized in FigTree 1.3.1. Multidimensional scaling (MDS) analysis was conducted in PLINK 1.07 based on either CNV genotypes generated in this study or SNP genotypes generated using the same bovine HapMap populations as described previously.

NimbleGen array CGH log2 ratios were calculated for each probe on a custom 2.1 million probe array for all animals. The reference animal, in all cases, was the Hereford cow used to generate the cattle reference assembly, L1 Dominey. All log2 ratio values that spanned the GAT, GLYAT, and KRTAP9-1 genes were averaged across the gene’s length for each animal individually. VST values contrasting taurine and indicine populations were calculated for these log2 ratios as previously described.
Heatmaps were generated using the estimated CN windows for each animal as described previously. The gplots (v 2.14.2) R package (http://cran.r-project.org/web/packages/gplots/index.html) was used to graph the CN values and generate heatmap representation of all lineage-specific gene duplications, deletions, and expansions identified in cattle breeds.

2.3. Gene analyses
Gene content of cattle CNVRs was assessed as previously described. We performed DAVID analysis to test whether the terms were under- or overrepresented in CNVRs after Bonferroni correction. We identified the lineage-specific or lineage-differential gene families using a heuristic approach with our 75 analysed individual animals. We divided the animals into breed, subspecies, and purpose groups as listed in Table 1 and used a weighted search algorithm to highlight CNVRs with a high tendency to exist solely within a specific group. The weighted search was accomplished as follows: for each CNVR we calculated a sum score that represented a hypothesis that the CNVR was unique to a specific breed/subspecies based on the animals that shared the CNVR. For each breed/subspecies/purpose group (G), we counted the number of animals (A) from G that shared the CNVR and imposed a penalty (P) for each animal that was not a member of the current G. The sets of G that were tested consisted of dairy, beef, Angus, Holstein, Limousin, Jersey, Romagnola, Nelore, Gir, Brahman, Taurus, and Indicus (membership was not mutually exclusive within the groups). The summed weight of A–P was calculated for each G, and if it exceeded a threshold of 3, it was selected as a putative subspecies/breed-specific or differential CNVR. We also employed a statistical method (VST) to identify copy number variable genes within our data set. Based on the gene CNs from each animal, we identified gene families that were stratified by subspecies differences using the statistic VST, as described above.

2.4. Haplotype network analysis
To explore the diversity of haplotypes and evolutionary relationships across populations, we retrieved the high-density SNP array data for these eight breeds generated by the Illumina BovineHD SNP Consortium as described previously. Haplotypes and their frequencies were calculated. To obtain reliable results, we employed an iterative scheme to perform inference with 10,000 iterations and 10,000 burn-ins, also we increased the number of iterations of the weighted search was accomplished as follows: for each CNVR we calculated a sum score that represented a hypothesis that the CNVR was unique to a specific breed/subspecies based on the animals that shared the CNVR. For each breed/subspecies/purpose group (G), we counted the number of animals (A) from G that shared the CNVR and imposed a penalty (P) for each animal that was not a member of the current G. The sets of G that were tested consisted of dairy, beef, Angus, Holstein, Limousin, Jersey, Romagnola, Nelore, Gir, Brahman, Taurus, and Indicus (membership was not mutually exclusive within the groups). The summed weight of A–P was calculated for each G, and if it exceeded a threshold of 3, it was selected as a putative subspecies/breed-specific or differential CNVR. We also employed a statistical method (VST) to identify copy number variable genes within our data set. Based on the gene CNs from each animal, we identified gene families that were stratified by subspecies differences using the statistic VST, as described above.

Table 1. Samples and sequence data sets

<table>
<thead>
<tr>
<th>Breed</th>
<th>Subspecies</th>
<th>Purpose</th>
<th>Animal count</th>
<th>Coverage range</th>
<th>CNV count</th>
<th>Average CNVs per animal</th>
</tr>
</thead>
<tbody>
<tr>
<td>Brahman (BRM)</td>
<td>Bos t. indicus</td>
<td>Beef</td>
<td>7</td>
<td>5–9×</td>
<td>3,836</td>
<td>548 (86)</td>
</tr>
<tr>
<td>Gir (GIR)</td>
<td>Bos t. indicus</td>
<td>Beef/dairy</td>
<td>6</td>
<td>5–14×</td>
<td>3,724</td>
<td>621 (30)</td>
</tr>
<tr>
<td>Nelore (NEL)</td>
<td>Bos t. indicus</td>
<td>Beef</td>
<td>8</td>
<td>6–20×</td>
<td>4,855</td>
<td>607 (38)</td>
</tr>
<tr>
<td>Angus (ANG)</td>
<td>Bos t. taurus</td>
<td>Beef</td>
<td>16</td>
<td>5–30×</td>
<td>11,657</td>
<td>729 (52)</td>
</tr>
<tr>
<td>Holstein (HOL)</td>
<td>Bos t. taurus</td>
<td>Dairy</td>
<td>22</td>
<td>4–20×</td>
<td>12,430</td>
<td>565 (80)</td>
</tr>
<tr>
<td>Jersey (JER)</td>
<td>Bos t. taurus</td>
<td>Dairy</td>
<td>6</td>
<td>4–13×</td>
<td>3,487</td>
<td>381 (46)</td>
</tr>
<tr>
<td>Limousin (LIM)</td>
<td>Bos t. taurus</td>
<td>Beef</td>
<td>6</td>
<td>5–10×</td>
<td>3,650</td>
<td>608 (48)</td>
</tr>
<tr>
<td>Romagnola (ROM)</td>
<td>Bos t. taurus</td>
<td>Beef/draft</td>
<td>4</td>
<td>6–10×</td>
<td>2,708</td>
<td>677 (20)</td>
</tr>
</tbody>
</table>

*Numbers in parentheses indicate 1 SD.

2.5. Data release
All array CGH data have been submitted to the gene expression omnibus (http://www.ncbi.nlm.nih.gov/geo/) under the accession number GSE62990. Raw data and population genetic and evolutionary analysis results are available upon request for research purposes. Raw reads were deposited under the SRA Bioproject PRJNA277147 in SRA (http://www.ncbi.nlm.nih.gov/sra/).

3. Results and discussion

3.1. CNV discovery and experimental validations
After carefully excluding samples with low coverage in our initial survey, we focused on the remaining 75 individuals in our final data set (see Table 1 and Supplementary Table S1 for sample information and sequence coverage). We identified CNVs using a sliding window approach based on the previously published MrsFAST-WSSD method. We discovered comparable average numbers of CNVs per individual across taurine (626.7) and indicine (591.2) cattle, suggesting our results based on the taurine reference assembly (UMD3.1) were not particularly biased against the indicine cattle. A full list of CNV calls (47,511) is presented in Supplementary Fig. S1 and Table S2. After merged across samples, these CNVs yielded 1,853 CNV regions (CNVR), which represent 875 Mb (3.1%) of the cattle genome (Supplementary Fig. S1 and Table S3). We then calculated absolute copy number values for 1 kb windows across the genome for each sequenced individual (see Materials and methods). As anticipated, the average normalized genome-wide copy number was 2.15 ± 0.1 for all copy number windows. We successfully performed 75 quantitative PCR (qPCR, Supplementary Table S4) and 25 array CGH experiments (Supplementary Fig. S2) to assess the false-positive discovery rate for our data set as previously described. Detailed experimental validations can be found in Supplementary Material online. In summary, CNV calls made with sequence data were strongly correlated with array CGH data (r2 = 0.761) and had an estimated 12% false-positive rate and a 19% false-negative rate based upon qPCR and array CGH, respectively.

We found that a large proportion of identified CNVRs (43.3%; 49.5 Mb) overlapped with the segmental duplication (SD) regions. We estimated pair-wise Spearman’s rank correlations (Supplementary Table S5) of 0.084 and 0.098 for indicine and taurine CNVs and SD regions (both P < 0.001), which were similar to the previously reported human results. A strong correlation of CNVs and SDs in cattle confirms that their formation mechanisms are mainly due to non-

Table 1. Samples and sequence data sets
allelic homologous recombination (NAHR).\textsuperscript{10,56} In the following analyses, we mainly focused on the characterization of the high-confidence CNVs from autosomes.

### 3.2. Population genetics of cattle CNVs

To investigate the population genetics of CNVs, we first identified the frequencies of CNVRs in our data set (Supplementary Fig. S3). The average CNVR had a frequency of 29.3\% (22 animals out of 75 total); however, the CNVR frequency tended towards a parabolic distribution with 835 CNVRs having a frequency $\leq 5\%$ and 189 CNVRs having a frequency $\geq 95\%$ in our data set (Supplementary Fig. S3). As expected, rare events were often observed in only one subspecies/breed, whereas common CNVs (frequency $>5\%$) were usually shared across subspecies/breeds.

To explore the population differentiation of CNVs between taurine and indicine cattle, we applied statistical measures of population differentiation using $V_{ST}$ to our dataset in three ways: (i) estimation of $V_{ST}$ for genome-wide 1 kb CN windows; (ii) clustering of the top 1\% of $V_{ST}$ values for 1 kb CN windows; and (iii) estimation of $V_{ST}$ using the average CN for annotated genes. Estimates of $V_{ST}$ for all genome-wide CN windows and all CNVRs revealed a number of outliers with levels of population differentiation suggestive of population-specific selective pressures (Fig. 1). Among these outliers were CNVs near AOX1, GAT/GLYAT, ASZ1, KRTAP9-1, and MCM4 (Fig. 1A).

We next selected the top 1\% diverse 1 kb windows ($n = 80$)—identified by our $V_{ST}$ calculations—from the distinct outlier CNVRs to perform CNV genotyping. The PAM algorithm is the most common implementation of k-medoid clustering, which is related to the k-means algorithm and the medoid shift algorithm.\textsuperscript{50} Previously, a PAM algorithm was used to cluster copy number derived from array CGH log$_2$ ratios into discrete CN genotypes.\textsuperscript{21} Similarly, we partitioned the copy numbers of each 1 kb window into three discrete value clusters, representing low, mid, and high ranges and then coded them as 0, 1, and 2 values within a matrix for subsequent genotyping (Supplementary Table S6).

Using CNV genotype calls from the top 1\% diverse 1 kb windows within CNVRs, we performed population clustering (Fig. 2). The proximity of an individual to each apex of the triangle indicates the proportion of that individual’s genome that is estimated to have ancestry from each of the three inferred ancestral populations. The clustering together of most indicine individuals (BRM, GIR, NEL) in the right bottom apex indicates the clear discrimination between indicine and taurine cattle. In contrast, taurine cattle are scattered along the opposing side with the exception of ROM in the centre. ANG individuals were clustered together in the upper apex, while the other taurine cattle (HOL, LMS, JER) were dispersed around the left bottom corner, suggesting a possible discrimination between beef and dairy cattle. This figure is available in black and white in print and in colour at DNA Research online.
while the other taurine cattle (HOL, LMS, JER) were dispersed around the left bottom corner, suggesting a distinction between different taurine breeds. It is possible that these two clusters differentiate between continental European breeds of cattle from UK breeds or beef breeds from dairy breeds; however, it is also possible that our selected CNV markers may be subject to cryptic founder effects within our ANG individuals. We suspect that the addition of African cattle breeds to this dataset will better resolve the taurine cluster by providing a third distinct lineage. Still, we note that the clustering results were structurally similar to our results obtained with high-density SNP data derived from the same bovine HapMap samples (Supplementary Fig. S4) and other recently published results.\textsuperscript{44} Based on the genotypes for these 80 loci (i.e. the top 1 kb windows) and using a neighbor-joining algorithm, we obtained a phylogenetic tree that generally agrees with the known cattle breed history (Supplementary Fig. S5). We also performed a MDS analysis based on CNV genotypes and compared it with the plot based on SNPs. Our plot confidently separated the indicine from the taurine cattle (Supplementary Fig. S6); however, the separation and clustering of the taurine cattle using CNVs were not superior to those based on SNPs, suggesting that CNV genotyping still has room for improvement.

Out of these 80 loci, 62 can be reliably assessed for their variable patterns and 54 of these loci, in turn, (87.10\%, 54/62) are located in or near tandem duplications (Supplementary Table S6). This estimate was consistent with our initial genome-wide results that 90\% of CNV in cattle are tandem duplications in contrast to human and other primates, which show a preponderance of interspersed duplications.\textsuperscript{56} This led us to speculate that while it is challenging to systematically genotype cattle duplication CNV events as shown by Genome STRiP results in human,\textsuperscript{57} our relatively high cattle CNV genotyping accuracy is likely due to the vast majority of cattle CNV being tandem duplicates. Large tandem repeats or duplication CNVs in cattle could behave similarly like human tandem macrosatellites and multicycle genes. For these tandem duplications, it is likely that we made reasonable approximations of CNV genotyping calls by simply clustering the normalized copy numbers, as shown traditionally for macrosatellites and microsatellites.\textsuperscript{58-60} Additionally, the tandem distribution pattern could contribute to the high LD at CNV loci as suggested previously,\textsuperscript{61} thus the majority of CNV genotype calls could better represent local alleles. Combining these two factors, it is not surprising that our CNV-based results generally agree with SNP-based results. Of course, this hypothesis certainly warrants more investigation using larger sample sizes and other mammals like mouse and dog to further validate and improve CNV genotyping approaches.

3.3. Gene analyses

We evaluated genes overlapped by cattle CNVs (Supplementary Table S7) and selected genes with known functions (Table 2). We observed an enrichment of CNVs intersecting with genes (P < 0.0001; Spearman’s rank sum correlation), consistent with reduced evolutionary constraints acting on functionally redundant gene categories. We next used DAVID to identify basic biological functional categories for 361 genes overlapped by our identified CNVRs.\textsuperscript{63} Like other mammals (human, mouse, and dog), statistically significant

<table>
<thead>
<tr>
<th>Gene name</th>
<th>Function</th>
<th>Gene UMD3.1 coordinates</th>
<th>$V_{ST}$</th>
<th>Identified$^a$</th>
</tr>
</thead>
<tbody>
<tr>
<td>AOX1</td>
<td>Detoxification</td>
<td>chr2:89517708-89589232</td>
<td>0.5094</td>
<td>Hou, Bickhart, and this study</td>
</tr>
<tr>
<td>ASZ1</td>
<td>Spermatogenesis</td>
<td>chr4:51294534-51370343</td>
<td>0.2090</td>
<td>Only this study</td>
</tr>
<tr>
<td>CAI</td>
<td>Carboxylic anhydrase</td>
<td>chr14:79520632-79530892</td>
<td>0.3270</td>
<td>Hou, Bickhart, and this study</td>
</tr>
<tr>
<td>CFH</td>
<td>Complement factor</td>
<td>chr16:5486704-6172566</td>
<td>0.0483</td>
<td>Hou, Bickhart, and this study</td>
</tr>
<tr>
<td>DDX21</td>
<td>Translation initiation</td>
<td>chr28:25376358-25399769</td>
<td>0.2375</td>
<td>Only this study</td>
</tr>
<tr>
<td>DENER</td>
<td>Translation initiation</td>
<td>chr29:7723699-7725004</td>
<td>0.3285</td>
<td>Only this study</td>
</tr>
<tr>
<td>FBXO16</td>
<td>Ubiquitin protein ligase</td>
<td>chr8:10095869-10128673</td>
<td>0.3538</td>
<td>Bickhart and this study</td>
</tr>
<tr>
<td>FZD3</td>
<td>Nervous system</td>
<td>chr8:10002971-10091175</td>
<td>0.3334</td>
<td>Bickhart and this study</td>
</tr>
<tr>
<td>GAL3T1</td>
<td>Glycoprotein</td>
<td>chr17:71660016-71678806</td>
<td>0.2435</td>
<td>Hou and this study</td>
</tr>
<tr>
<td>GAMT</td>
<td>Detoxification</td>
<td>chr15:83472190-83493607</td>
<td>0.4336</td>
<td>Liu, Hou, Bickhart, and this study</td>
</tr>
<tr>
<td>GLYAT1</td>
<td>Detoxification</td>
<td>chr15:83455512-83469280</td>
<td>0.4083</td>
<td>Liu, Hou, Bickhart, and this study</td>
</tr>
<tr>
<td>GLYATL2</td>
<td>Biological oxidation</td>
<td>chr15:83508339-83515102</td>
<td>0.2257</td>
<td>Bickhart and this study</td>
</tr>
<tr>
<td>KRTAP9-1</td>
<td>Keratin family</td>
<td>chr19:42101853-42103421</td>
<td>0.4578</td>
<td>Bickhart and this study</td>
</tr>
<tr>
<td>LMBR2D</td>
<td>Function unknown</td>
<td>chr20:38116509-38163145</td>
<td>0.2573</td>
<td>Only this study</td>
</tr>
<tr>
<td>PGR</td>
<td>Progesterone receptor</td>
<td>chr15:8207682-8222806</td>
<td>0.0103</td>
<td>Only this study</td>
</tr>
<tr>
<td>PNPLA2</td>
<td>Adipose tissue regulation</td>
<td>chr29:50742384-50747161</td>
<td>0.0000</td>
<td>Only this study</td>
</tr>
<tr>
<td>PRG3</td>
<td>Carbohydrate binding</td>
<td>chr15:81920283-81926082</td>
<td>0.2041</td>
<td>Liu, Hou, and Bickhart, and this study</td>
</tr>
<tr>
<td>RAI1G/ULBP17</td>
<td>MHC class I related</td>
<td>chr9:88231932-88402262</td>
<td>0.0000</td>
<td>Liu, Hou, and Bickhart, and this study</td>
</tr>
<tr>
<td>RICTOR</td>
<td>Cell growth</td>
<td>chr20:35376523-35314733</td>
<td>0.1048</td>
<td>Only this study</td>
</tr>
<tr>
<td>SEC3A</td>
<td>Vesicle transport</td>
<td>chr21:49489514-49553507</td>
<td>0.0305</td>
<td>Only this study</td>
</tr>
<tr>
<td>SERPINB4</td>
<td>Protease inhibitor</td>
<td>chr24:62364701-62371668</td>
<td>0.2418</td>
<td>Liu, Hou, and Bickhart, and this study</td>
</tr>
<tr>
<td>SUB1</td>
<td>Transcriptional activation</td>
<td>chr20:41122022-41143914</td>
<td>0.2282</td>
<td>Only this study</td>
</tr>
<tr>
<td>TMED2</td>
<td>Secretory vesicle transport</td>
<td>chr17:54330420-54338333</td>
<td>0.0000</td>
<td>Only this study</td>
</tr>
<tr>
<td>UFH1</td>
<td>Ubiquitin</td>
<td>chr6:71051155-71053333</td>
<td>0.5121</td>
<td>Only this study</td>
</tr>
<tr>
<td>ZNF280B</td>
<td>Negative regulation of p53</td>
<td>chr17:51251538-51262528</td>
<td>0.2638</td>
<td>Liu, Hou, and this study</td>
</tr>
</tbody>
</table>

$^a$ $V_{ST}$ was calculated from the comparison between the taurine and indicine individuals.

$^b$Liu, Hou, and Bickhart: we focused on the comparisons with the published CNV results based on the same bovine HapMap samples using array CGH,\textsuperscript{10} BovineHD SNP array,\textsuperscript{15} and individual NGS.\textsuperscript{16} respectively.
overrepresentations were observed for multiple categories including chromosome maintenance, immunity and cytoskeleton components (Supplementary Table S8). We then studied how variable genes were distributed across subspecies/breeds using either a heuristic approach based on CNV presence/absence or gene CN per individual ($V_{ST}$).

3.4. Lineage-specific CNV genes based on a heuristic approach

We first identified lineage specific, copy number variable genes (CNV genes) using a heuristic approach (see Materials and methods). Dairy cattle-specific CNVs tended to be present at low frequencies in our 28 dairy cattle, and they manifested as small copy number changes of affected genes. Several of these dairy-specific CNVRs were found to intersect genes related to cellular growth and development pathways, including $RICTOR$ (rapamycin-insensitive companion of mTOR)$^{64}$ and $TMED2$. We also identified several lipid metabolism genes that overlapped CNVRs exclusive to beef cattle (over 40 samples). Within our Angus data set, we discovered six animals that had a predicted heterozygous duplication of $PNPLA3$ (the Patatin-like phospholipase domain-containing protein 3). This gene is expressed in adipose tissue and liver, and is associated with the $de$ $novo$ synthesis of fatty acids.$^{65}$ In indicine beef cattle (Nelore, Brahman, and Gir), we also detected a duplication in a predicted Ensembl gene (ENSBTA00000043749) containing functional domains related to lipid metabolism.

3.5. Gene family expansion, diversity, and evolution

We identified copy number variable genes among the different subspecies/breeds using $V_{ST}$ statistics. We defined highly stratified genes as genes having $V_{ST}$ values $>0.2$. CN plots for these stratified genes showed clear differences in the average CN value for taurine and indicine animals (Fig. 3A). Based on $V_{ST}$ values, ZNF280B, FBXO16, $KRTAP9-1$, $MCM4$, $SERPINB4$, $CA1$, $FZD3$, GLAYAT, GAT, MANBA, and $DENR$ were the most stratified genes. To provide orthogonal experimental support for the sequence-based $V_{ST}$ results, we also retrieved log2 ratios from array CGH data for the same animals. Representative results are shown in Supplementary Fig. S8 for the GAT, GLAYAT, and KRTAP9-1 genes, further confirming the sequence-based $V_{ST}$ results.

CN stratified genes tended to be immune system related, which is expected given the different environmental challenges in the history of evolution of taurine and indicine cattle. One of these stratified CNVs represents a significantly higher duplication of the $KRTAP9-1$ gene in taurine cattle, which is a paralog of $KRTAP9-2$ that was previously reported to likely be involved in indicine tick resistance (Fig. 3B). A duplication of the $MCM4$ gene (Fig. 3C) was found in indicine cattle compared with taurine cattle. We also confirmed several other gene families appearing to be copy number variable, including lysozyme, defensin, and unique long binding protein ($ULBP$) families and the major histocompatibility complex (MHC).

Discoveries of high-frequency gene duplications suggest that the affected gene families are currently expanding in a ‘gene family birth and death’ model as described by Nei and Rooney.$^{67}$ Such multicopy genes, if present in a sufficiently large proportion of the population, can be thought of as signs of diversifying selection or selection by overdominance.$^{68}$ One example of this can be found in the olfactory receptor (OR) gene family, which has several member genes that detect odorant molecules through combinatorial binding across other paralogous family members.$^{69}$ Therefore, the duplication and subsequent

![Figure 3](http://dx.doi.org/10.1093/dnares/dsw204) Cattle gene family copy number diversity and evolution. The genes most stratified by copy number on the basis of $V_{ST}$ analysis of taurine and indicine cattle (A). The most copy number variable genes in both taurine and indicine subspecies (legend insets denote group colors) tended to be immune system related genes. Histograms showing the distributions of copy numbers among the unrelated individuals in each group are plotted for the $KRTAP9-1$ gene (B) and the $MCM4$ gene (C). X-axis values indicate copy number and Y-axis values indicate sample count. Individual copy number values for each gene can be found in Supplementary Table S7. This figure is available in black and white in print and in colour at DNA Research online.
mutation of OR gene members allow for a greater range of odorant detection for a host organism. Indeed, out of 134 annotated OR genes, we have identified 31 (23.1%) individual genes that have predicted duplications in our data set.

We have detected several additional gene families that appear to be subject to a high degree of duplication in our data set, and these families likely represent classes of genes that are in the processes of sub-functionalization and neofunctionalization in cattle. They include a cluster of prolactin-related protein family (PRP) genes that appears to be duplicated in 96% (74/75) animals. It was previously discovered using the BovineSNP50 array; however, we refined the event from 2.4 down to 0.7 megabases in size. Another locus containing several pregnancy-associated glycoprotein (PAG) family members was found to be duplicated in all animals within CNVR 1717.

3.6. Haplotype network analyses near selected multicopy genes

It is important to note that CNVs in some loci may have different alleles. Earlier results also suggest that the diversity of a subset of multicopy genes like human OR genes may have been maintained by balancing selection, in the form of overdominance. For example, a 660 kb deletion with antagonistic effects on fertility and milk production was recently found at high frequency in Nordic Red cattle, providing evidence for balancing selection of CNVs in livestock. To investigate the potential effect of overdominance on selection and evolution of multicopy genes, we further investigated haplotype evolution pattern using the BovineHD SNP array. We obtained 11 haplotypes within the 50.3 kb haplotype region near the GLYAT/GAT locus (Fig. 4A). The most common haplotype, H1 (with frequency of 70.06%), was mainly found in taurine cattle (HOL, ANG, JER, and LMS) and only minor portions were found in indicine cattle (ROM, BRM, GIR, and NEL) (Fig. 4A). H2 (with frequency of 10.56%) included a large proportion of taurine cattle (ANG, LMS, and ROM). We also observed two haplotypes exclusive to indicine cattle with a combined frequency of 10.54%: H3 and H4. Altogether, this pattern indicated that separate haplotypes were clustered only for the indicine cattle (BRM, GIR, and NEL), while other common haplotypes were identified for taurine cattle (HOL, ANG, LMS, JER, and ROM). BRM and ROM were the only exceptions, as they were often associated with both taurine and indicine cattle. This was not unexpected, as it mirrors the complex ancestral backgrounds of these two breeds, since BRM cattle are a known indicine breed with taurine influence and ROM share distinctive genetic ancestry with indicine cattle.

3.7. The impacts of the reference genome assembly

Different versions of cattle reference genome assemblies (Btau_4.0 and UMD3.1) have different RefSeq gene annotations, particularly in CNV regions of the cattle genome. For example, the CATHL4 gene, for which we previously reported copy number change, was located on chrUn of UMD3.1. Since our RD method uses a window approach that relies on large genomic segments, this prevented us from assessing the copy number status of CATHL4 on chrUn. However, we did find CATHL4 was copy number variable in this study. Similarly, we were unable to assess the copy number of KRTAP9-2, however, we detected copy number changes for one of its paralogs, KRTAP9-1, in this study.

3.8. Limitations and future directions

Similar to SNP and microsatellite, CNV distribution within and among populations seems to be shaped by mutation, recombination, gene conversion, selection, and demographic history. However, CNV genetic markers may not be currently compatible with current population analyses, because CNVs violate the classical population genetics assumptions based on the infinite allele model and the infinite site model for SNP. Compared with SNPs, limitations of CNVs as markers were observed in this study probably due to their distinct mutation mechanisms, high mutation rates, heterogeneities among loci, and uncertainties related to allele calling. Similar observations were also reported for microsatellite—primarily due to similar limitations of detection and variability. For example, although NAHR is believed to be responsible for most of large duplication CNVs in cattle, inference, and predictions on the forces influencing populations require the modelling of the mutational process generating CNV. We currently lack such a model for the large duplications present in cattle. This is also compounded by the fact that homoplasy
caused by recurrent events is expected to occur relatively often for CNV compared with SNP because of their high mutation rates.

Despite the aforementioned limitations, this study represents one of the first attempts to genotype CNVs within large, diverse cattle populations using sequence data. Although beyond the scope of this study, a comparison with human-centric CNV genotyping methods using cattle sequence data will provide a useful contrast in approaches. Our results provide a new glimpse into the diversity of selective pressures during cattle speciation. We confirmed that cattle are strikingly diverse, despite relatively low estimated current population sizes for several taurine cattle as shown previously.37 Our population-genetic analyses based on CNVs reveal the population structures of these taurine and indicine cattle and uncover hundreds of CNVs showing elevated population differentiation near important functional genes. We highlighted several subspecies specific or differential CNV gene overlaps that are likely subject to subfunctionalization and neofunctionalization. We also identified key regions of the cattle genome that are subject to variation and reported several potential genes affecting productive traits. These discoveries provide a basis for future efforts to genotype and track large CNVs in cattle. More sequencing data from the 1000 Bull Genomes Project13 or the analysis of additional outlier groups (e.g. African cattle breeds) will help to validate and refine the link between genomic copy number in these regions or different alleles with production and health traits.

Authors’ contributions

Acknowledgements
We thank members of the Illumina BovineHD SNP Consortium for sharing their samples and data. We would thank the Council on Dairy Cattle Breeding (Reynoldsburg, OH) for access to animal phenotypic information, and we thank the Cooperative Dairy DNA Repository (Beltsville, MD) for access to semen stocks used to generate Illumina DNA sequence libraries. We also thank T. Brown, R. Anderson, and A. Dimtchev for technical assistance. Mention of trade names or commercial products in this article is solely for the purpose of providing specific information and does not imply recommendation or endorsement by the US Department of Agriculture.

Supplementary data
Supplementary data are available at www.dnaresearch.oxfordjournals.org.

Funding
This work was supported in part by Agriculture and Food Research Initiative (AFRI) competitive grant No. 2011-67015-30183 from the USDA National Institute of Food and Agriculture (NIFA) Animal Genome Program. J.F.T. and C.P.V.T. were supported by AFRI competitive grant No. 2009-65205-05635 from the USDA NIFA. J.F.T. and R.D.S. were further supported by AFRI competitive grants No. 2011-68004-30367, 2011-68004-30214 and 2013-68004-20364 from the USDA NIFA Animal Genome Program. Funding to pay the Open Access publication charges for this article was provided by the USDA National Institute of Food and Agriculture (NIFA) Animal Genome Program.

Conflict of interest statement
T.S.S. is an employee of Acceligen Inc. of Animal Agriculture Subsidiary of Recombinetics, Inc. All other authors declare no potential conflict of interest.

References


